Negotiating AI Safety Globally

More than 1300 employees of US frontier AI companies—including many of their most prominent scientists—signed an open letter warning that automated AI research could soon accelerate their models' capabilities faster than humans' ability to control them. They call for an international effort to develop the technical means to address this risk and "pace" the frontier of AI development.

They're right.

While several members of Congress have introduced bills to set sensible domestic rules for AI safety, such measures remain insufficient when developers in the U.S. and the rest of the world increasingly rely upon open-source Chinese models beyond the reach of American law. Our ability to forge consensus on meaningful domestic regulation will continue to remain undermined so long as we cannot adequately respond to the question, "How can we make Americans any safer with U.S. regulation that ties the hands of American AI labs, but cannot reach hundreds of millions of Chinese open-source models that produce nearly as powerful (and at least as dangerous) AI agents?"

To halt a "race to the bottom" on AI safety, we need to establish a shared framework that ensures developers in both the US and China adhere to the same standards and constraints, paired with strong verification measures. Whatever safety rules and standards the U.S. and China can settle upon, the rest of the world will likely follow—or suffer technological exclusion.

This proposal carries no delusions about this Administration's current willingness to lean in to safety regulation, or to any negotiation with Beijing on the topic. Regardless of when the Executive Branch chooses to engage, however, much work must get done. The table must be set with the extensive technical work of experts in the field of AI who can find common ground around metrics, standards, protocols, and the like. Industry, civil society, and academia should begin working with their foreign counterparts now—starting with those in China—to develop the technical standards and capabilities necessary for future bilateral or multilateral agreements, whenever that policy window opens.

Outline of Proposed Legislation

Section 1: The Secretary of State, in consultation with the Secretary of Commerce, Secretary of Treasury, Secretary of Defense, Director of National Intelligence, and others the Secretary deems appropriate, shall seek to commence negotiations with the government of China to:

- Identify shared priorities for mitigating frontier AI safety and catastrophic risk, including establishing agreed-upon standards, model testing and evaluation, compliance verification, third-party auditing, capability pacing, transparency in pre-deployment evaluation, incident notification, and any other risk management measures prioritized by the Secretary.
- Focus discussions on risk mitigation measures to address catastrophic risk, namely (a) alignment and loss of control, (b) chemical, biological, radiological, nuclear weapons (CBRN), and (c) cybersecurity.
- Negotiate with the objective of reaching a treaty or other binding agreement between both nations to regulate frontier AI developers in each country to adhere to shared standards to address these risks, with clear mechanisms to verify compliance.
- Establish initial negotiations as bilateral, but enable subsequent expansion of commitments based upon US-China agreements, including the establishment of an international AI safety body.

- Establish a hotline between the US and Chinese governments, modeled on the "red phone" between Washington and Moscow, for direct communication on AI risk.

Section 2: CAISI shall establish an initiative composed of Al safety experts from industry, academia, and civil society to engage with their counterparts in other countries to promote development and adoption of standards related to Al safety. In its engagement with Al safety institutes in other countries or through US participation in international standards bodies such as the International Organization for Standardization, CAISI shall prioritize

- the development and harmonization of Al safety testing and evaluation metrics,
- means to verify compliance, and
- other efforts necessary to establish common understanding to facilitate agreement on bilateral safety measures described in section 1.

An amount of $100 million shall be authorized to support the operations of CAISI in this endeavor.

Section 3: Until the conclusion of a treaty or other binding agreement described in Section 1, Al industry and safety experts from universities, think tanks, and industry Al labs may communicate directly with peer experts within the People's Republic of China without violating federal laws, including the Sherman Act §1 and other antitrust laws, the Economic Espionage Act (18 U.S.C. §§1831-1832), the Espionage Act provisions, 18 U.S.C. §951, the Foreign Agents Registration Act, the International Emergency Economic Powers Act, the Export Control Reform Act, and Export Administration Regulations, so long as those communications meet all of the following conditions:

1. Registration and U.S. Participation. (1) U.S. participants must register with the U.S. Secretary of State, in a manner determined by the Secretary, no fewer than 14 days prior to any such communication, and (2) the Secretary may designate a member of government who must be allowed to observe any communication, including electronic and verbal communications.
2. Subject matter. Communication must be strictly limited to developing common understanding regarding testing and evaluation standards, protocols, metrics, and methods necessary to facilitate agreement between the U.S. and China to mitigate Al catastrophic safety risks, as described in Section 1.
3. Limitations. U.S. persons may not provide any information during such communications that (a) facilitates the distillation of U.S. large language models; (b) enables the release of intellectual property possessed by a U.S. person or entity, (c) shares non-public information about U.S. cybersecurity or national security vulnerabilities that could facilitate exploitation of those vulnerabilities.

Section 4: Requires both the Secretary of State and the Secretary of Commerce to report to Congress with monthly updates on their progress.

Section 5: Separately, this Act directs that the Secretary of Commerce, in consultation with other agencies, shall seek to commence negotiations with the government of China regarding measures to ensure protection of U.S. intellectual property compromised by model distillation.